Analysis of Linear Convergence of a (1 + 1)-ES with Augmented Lagrangian Constraint Handling

Asma Atamna, Anne Auger, Nikolaus Hansen

Analysis of Linear Convergence of a $(1+1)$-ES with Augmented Lagrangian Constraint Handling

Asma Atamna
Inria
Centre Saclay–Île-de-France
LRI, Université Paris-Saclay
atamna@lri.fr

Anne Auger
Inria
Centre Saclay–Île-de-France
LRI, Université Paris-Saclay
auger@lri.fr

Nikolaus Hansen
Inria
Centre Saclay–Île-de-France
LRI, Université Paris-Saclay
hansen@lri.fr

ABSTRACT

We address the question of linear convergence of evolution strategies on constrained optimization problems. In particular, we analyze a $(1+1)$-ES with an augmented Lagrangian constraint handling approach on functions defined on a continuous domain, subject to a single linear inequality constraint. We identify a class of functions for which it is possible to construct a homogeneous Markov chain whose stability implies linear convergence. This class includes all functions such that the augmented Lagrangian of the problem, centered with respect to its value at the optimum and the corresponding Lagrange multiplier, is positive homogeneous of degree 2 (thus including convex quadratic functions as a particular case). The stability of the constructed Markov chain is empirically investigated on the sphere function and on a moderately ill-conditioned ellipsoid function.

Categories and Subject Descriptors

I.2.8 [**Problem Solving, Control Methods, and Search**]; G.1.6 [**Optimization**]: Constrained Optimization

Keywords

Augmented Lagrangian, constrained optimization, evolution strategies, Markov chains

1. INTRODUCTION

Linear convergence is central in the study of evolution strategies (ESs). Ideally, we want an ES to converge linearly on the widest possible range of optimization problems. As illustrated in [5] for unconstrained optimization, linear convergence can be derived on scaling-invariant functions by exploiting invariance properties of the algorithm at hand on this class of functions: invariance allows to exhibit a Markov chain whose stability leads to linear convergence. In this context, stability is defined as positivity and Harris-recurrence, and is usually obtained by proving φ-irreducibility, aperiodicity, and the existence of a drift function on a small set [10, 5]. Linear convergence then follows from the application of a Law of Large Numbers (LLN). To see how this methodology is applied in practice, one can refer to [4] where linear convergence is proven for the $(1,\lambda)$-ES with self-adaptation on the sphere function, or [6] where the authors show linear convergence of the $(1+1)$-ES with $1/5$th success rule on the class of positive homogeneous functions. Stability is generally difficult to prove "manually". In an attempt to reduce this difficulty, the authors in [7] propose a set of sufficient conditions for a Markov chain to be irreducible and aperiodic.

Linear convergence is also desired on constrained optimization problems [3]. However, little is known about how it can be achieved. Most theoretical works on ESs in the constrained case deal with linear problems with a single linear constraint, as in [2, 1] where the single-step behavior of the $(1+1)$-ES and the $(1,\lambda)$-σSA-ES is analyzed on the linear function with a single linear constraint. In [3], linearly constrained convex quadratic problems are studied for the first time. The authors present an inequality constraint handling method for the $(1+1)$-ES based on augmented Lagrangian and analyze the single-step behavior of the algorithm on the sphere function with one linear inequality constraint. Based on this analysis, they design an update rule for the penalty parameter of the augmented Lagrangian so that the algorithm is empirically observed to converge linearly on sphere and moderately ill-conditioned ellipsoid problems.

In this work, we go one step further into understanding theoretically how linear convergence can be achieved for ESs implementing an augmented Lagrangian constraint handling approach. We introduce a variant of the algorithm presented in [3] and analyze its behavior on the problem of minimizing a function defined on a continuous domain, subject to a single linear inequality constraint. We show that for objective functions such that the corresponding augmented Lagrangian minus its value at the optimum and the corresponding Lagrange multiplier is positive homogeneous of degree 2, one can construct a homogeneous Markov chain and prove linear convergence assuming its stability. Similarly to the unconstrained case, invariance is a key element for constructing the Markov chain. However, invariance alone is not sufficient and another key element is how the parameters of the augmented Lagrangian are updated. Assuming the Markov chain is stable, we prove linear convergence of the solution at a given iteration towards the optimum of the problem, as well as linear convergence of both the Lagrange factor and the step-size towards the Lagrange multiplier associated to the optimum and zero respectively. Then, we empirically investigate the stability of the constructed Markov chain.

The rest of this paper is organized as follows: we formally define the optimization problem we consider in Section 2 and discuss the augmented Lagrangian method in Section 3. We present our algorithm and discuss its invariance properties in Section 4. In Section 5, we present the Markov chain and prove linear convergence

GECCO'16, July 20-24, 2016, Denver, Colorado, USA.

assuming its stability. We present our empirical results in Section 6 and conclude with a discussion on the main result of this paper in Section 7.

1.1 Notations

We define here all the notations which are not explicitly presented in the paper. We denote $\mathbb{R}^+$ the set of positive real numbers and $\mathbb{R}^+_>$ the set of strictly positive real numbers. $\mathbf{x} \in \mathbb{R}^n$ is a column vector, $\mathbf{x}^T$ is its transpose, and $\mathbf{0} \in \mathbb{R}^n$ is the zero vector. $\|\mathbf{x}\|$ denotes the Euclidean norm of $\mathbf{x}$, $\sim$ equality in distribution, and $\circ$ the function composition operator. The notation $(1+1)$ represents the "one-plus-one" selection scheme. $\mathbf{I}_{n\times n} \in R^{n\times n}$ denotes the identity matrix and $\mathcal{N}(\mathbf{0}, \mathbf{I}_{n\times n})$ the multivariate standard normal distribution. $[\mathbf{x}]_i$ is the ith component of vector $\mathbf{x}$ and $[\mathbf{M}]_{ij}$ is the element in the ith row and jth column of matrix $\mathbf{M}$. The derivative with respect to $\mathbf{x}$ is denoted $\nabla_\mathbf{x}$ and the expectation of a random variable $X \sim \pi$ is denoted E_π. Finally, $\mathbf{1}_{\{A\}}$ returns 1 if A is true and 0 otherwise.

2. OPTIMIZATION PROBLEM

We consider the problem of minimizing a function $f : \mathbb{R}^n \to \mathbb{R}$, n is the dimension of the search space, subject to one linear constraint $g(\mathbf{x}) \leq 0$, where $g : \mathbb{R}^n \to \mathbb{R}$. More formally, we write

$$\min_{\mathbf{x}} f(\mathbf{x}) \;\text{ subject to }\; g(\mathbf{x}) = \mathrm{b}^T\mathbf{x} + c \leq 0 \;, \tag{1}$$

where $\mathrm{b} \in \mathbb{R}^n$ and $c \in \mathbb{R}$. We assume the problem to admit a unique global minimum $\mathbf{x}_{\text{opt}}$ and the constraint to be active at $\mathbf{x}_{\text{opt}}$, that is, $g(\mathbf{x}_{\text{opt}}) = 0$.

We consider throughout this paper an ES based on the so-called augmented Lagrangian approach for handling constraints to seek the minimum of this problem. In the next section, we give general notions about the augmented Lagrangian approach.

Since we consider only minimization problems, we will sometimes refer to the minimum as the optimum in the rest of this paper.

3. AUGMENTED LAGRANGIAN APPROACH

The augmented Lagrangian approach for handling constraints is a combination of the Karush-Kuhn-Tucker (KKT) and penalty function methods. It was introduced for the first time in [8] and [12]. The KKT method defines first-order optimality conditions, referred to as KKT conditions. It introduces the Lagrangian $\mathcal{L} : \mathbb{R}^{n+1} \to \mathbb{R}$ defined as

$$\mathcal{L}(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda g(\mathbf{x}) \;, \tag{2}$$

$\mathbf{x} \in \mathbb{R}^n$, $\lambda \in \mathbb{R}$, for an objective function f subject to one inequality constraint $g(\mathbf{x}) \leq 0$. Given some regularity conditions - or constraint qualifications - are satisfied, if $\mathbf{x}^* \in \mathbb{R}^n$ is a local minimum of the constrained problem such that f and g are continuously differentiable at $\mathbf{x}^*$, then there exists a non-negative constant λ^*, called the Lagrange multiplier, such that $\nabla_\mathbf{x}\mathcal{L}(\mathbf{x}^*, \lambda^*) = \mathbf{0}$, that is, $\mathbf{x}^*$ is a stationary point for $\mathcal{L}(\mathbf{x}, \lambda^*)$ (stationarity KKT condition). Put differently, given the "right" λ, the optimum of the constrained problem is a stationary point of the Lagrangian.

Considering the optimization problem in (1) and ellipsoid functions $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x}$, where $\mathbf{H} \in \mathbb{R}^{n\times n}$ is a diagonal matrix with diagonal elements $[\mathbf{H}]_{ii} = \alpha^{\frac{i-1}{n-1}}$, $\alpha > 0$, KKT conditions are satisfied for the unique minimum of the problem

$$\mathbf{x}_{\text{opt}} = -\frac{c}{\mathrm{b}^T\mathbf{H}^{-1}\mathrm{b}}\mathbf{H}^{-1}\mathrm{b} \tag{3}$$

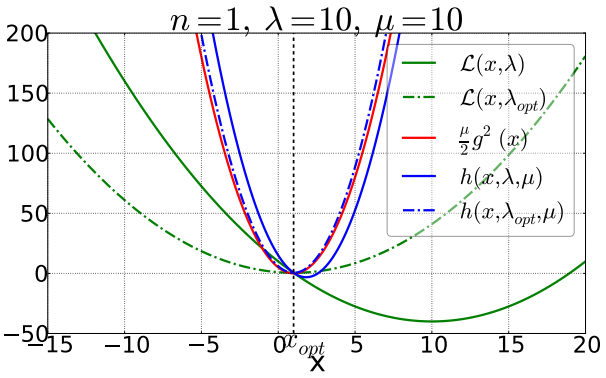

Figure 1: Graphs of $\mathcal{L}(x, \lambda)$ (green), $\mathcal{L}(x, \lambda_{\text{opt}})$ (dashed green), $\frac{\mu}{2}g^2(x)$ (red), $h(x, \lambda, \mu)$ (blue), and $h(x, \lambda_{\text{opt}}, \mu)$ (dashed blue) for $\lambda = 10$ and $\mu = 10$ in $n = 1$. $f(x) = \frac{1}{2}x^2$ and $g(x) = -x + 1$. $x_{\text{opt}} = 1$ and $\lambda_{\text{opt}} = 1$.

and the Lagrange multiplier

$$\lambda_{\text{opt}} = \frac{c}{\mathrm{b}^T\mathbf{H}^{-1}\mathrm{b}} \;. \tag{4}$$

In augmented Lagrangian approaches, the Lagrangian in (2) is combined with a penalty term, resulting in the augmented Lagrangian function h. The motivation for using the augmented Lagrangian is to overcome the shortcomings of quadratic penalty function methods, where the penalty factor needs to tend to infinity to achieve convergence [11]. This results in an ill-conditioned problem.

Different formulations of the augmented Lagrangian are possible depending on the optimization problem at hand. A broader discussion on augmented Lagrangians is provided in [11]. In our optimization problem, the constraint is active at the optimum $\mathbf{x}_{\text{opt}}$. Therefore, we use the following augmented Lagrangian

$$h(\mathbf{x}, \lambda, \mu) = f(\mathbf{x}) + \lambda g(\mathbf{x}) + \frac{\mu}{2}g^2(\mathbf{x}) \;, \tag{5}$$

where a quadratic penalty term $\frac{\mu}{2}g^2(\mathbf{x})$ is added to penalize points lying outside the boundary of the constraint, μ is a positive penalty factor. At each iteration, h is minimized with respect to $\mathbf{x}$. The parameters λ and μ are updated in such a way that λ approaches the Lagrange multiplier while μ guides the search towards solutions on the constraint boundary. Note that the optimum $\mathbf{x}_{\text{opt}}$ (which is also a KKT point) satisfies $\nabla_\mathbf{x} h(\mathbf{x}_{\text{opt}}, \lambda_{\text{opt}}, \mu) = \mathbf{0}$, for all $\mu \in \mathbb{R}^+_>$, where λ_{opt} is the Lagrange multiplier associated to $\mathbf{x}_{\text{opt}}$.

Figure 1 shows graphs of the penalty function $\frac{\mu}{2}g^2$, the Lagrangian $\mathcal{L}$, and the augmented Lagrangian h associated to the sphere function $f(x) = \frac{1}{2}x^2$ in dimension $n = 1$, with $g(x) = -x + 1$. KKT conditions are satisfied for the optimum $x_{\text{opt}} = 1$ and the Lagrange multiplier $\lambda_{\text{opt}} = 1$. $\mathcal{L}$ and h are plotted for $\lambda = 10, \lambda_{\text{opt}}$, and $\mu = 10$. For $\lambda = \lambda_{\text{opt}}$, the minimum of both $\mathcal{L}$ and h (dashed green and blue graphs) correspond to x_{opt}. However, for $\lambda = 10$, the minimum of $\mathcal{L}$ is different (green graph). By adding a penalty term (red graph) to the Lagrangian, the minimum of the augmented Lagrangian (blue graph) moves closer to x_{opt}.

Remark 1. The augmented Lagrangian in (5) is designed for the very specific case of an active constraint ($g(\mathbf{x}_{\text{opt}}) = 0$). This choice is motivated by theoretical considerations–mainly the construction of a homogeneous Markov chain. Note that for problems where $\mathbf{x}_{\text{opt}}$ is inside the feasible domain, i.e. $g(\mathbf{x}_{\text{opt}}) \neq 0$ and $\lambda_{\text{opt}} = 0$, $\nabla_\mathbf{x} h(\mathbf{x}_{\text{opt}}, \lambda_{\text{opt}}, \mu) \neq \mathbf{0}$. Hence, in practice where such an informa-

tion about the optimum is not provided, the augmented Lagrangian used in [3] is the appropriate choice.

4. ALGORITHM

In this section, we present a $(1+1)$-ES for solving the optimization problem described in (1), based on the augmented Lagrangian approach described above. The algorithm, summarized in Algorithm 1, iteratively minimizes the augmented Lagrangian function h (5) and adapts the Lagrange and penalty factors λ and μ. It is largely based on the $(1+1)$-ES presented in [3]. Indeed, we use the same update for μ. For λ, however, we modify the update used in [3]. This modification indeed seems to be necessary to be able to exhibit a Markov chain whose stability leads to linear convergence.

Algorithm 1 is a randomized adaptive algorithm. A general randomized adaptive algorithm optimizing a function $f : \mathbb{R}^n \to \mathbb{R}$ subject to a constraint $g(\mathbf{x}) \leq 0$, $g : \mathbb{R}^n \to \mathbb{R}$, is a sequence $(\mathbf{s}_t)_{t\in\mathbb{N}}$ of states, where $\mathbf{s}_t \in \Omega$ is the state of the algorithm at iteration t. The sequence is defined recursively as

$$\mathbf{s}_{t+1} = \mathcal{F}^{(f,g)}(\mathbf{s}_t, \mathbf{U}_{t+1}) \ , \tag{6}$$

where $\mathcal{F}^{(f,g)} : \Omega \times \mathbb{U}^p \to \Omega$ is the transition function of the algorithm and $(\mathbf{U}_{t+1})_{t\in\mathbb{N}}$ is a sequence of independent identically distributed (i.i.d.) random vectors $\mathbf{U}_t \in \mathbb{U}^p$ [5]. For Algorithm 1, the state at iteration t is given by $(\mathbf{X}_t, \sigma_t, \lambda_t, \mu_t)$ where $\mathbf{X}_t \in \mathbb{R}^n$ is the current solution, $\sigma_t \in \mathbb{R}^+$ is the current step-size, $\lambda_t \in \mathbb{R}$ is the current Lagrange factor, and $\mu_t \in \mathbb{R}^+_>$ is the current penalty factor. In fact, Algorithm 1 is based on the $(1+1)$-ES with $1/5$th success rule designed for unconstrained optimization where two additional state variables, λ_t and μ_t, are added to the original state $(\mathbf{X}_t, \sigma_t)$. Indeed, the fitness (the augmented Lagrangian here) in the constrained case is dynamic and is determined by λ_t and μ_t, which are adapted besides $\mathbf{X}_t$ and σ_t.

Given the current state $(\mathbf{X}_t, \sigma_t, \lambda_t, \mu_t)$, a standard normally distributed vector $\mathbf{Z}_{t+1} \in \mathbb{R}^n$ is sampled. It is then multiplied by the step-size σ_t and added to the current solution $\mathbf{X}_t$ to create the first candidate solution $\mathbf{X}^1_{t+1}$, according to Line 3 of Algorithm 1. The second candidate solution is $\mathbf{X}_t$. $\mathbf{X}^1_{t+1}$ and $\mathbf{X}_t$ are then ranked according to their fitness values, where the fitness at iteration t is defined by $h(\mathbf{x}, \lambda_t, \mu_t)$ for a given $\mathbf{x} \in \mathbb{R}^n$. The best point becomes the solution $\mathbf{X}_{t+1}$ at the next iteration. This is done in Lines 4 and 8 by computing the fitness difference Δh.

The step-size σ_t is adapted with the $1/5$th success rule [9]. It is multiplied by $2^{1/n}$ when $\mathbf{X}^1_{t+1}$ is better than $\mathbf{X}_t$ fitness-wise (Line 9) and by $2^{-1/(4n)}$ otherwise (Line 11). The idea behind this update is to increase (respectively decrease) the step-size if the success probability is larger (respectively smaller) than $1/5$.

The Lagrange factor λ_t is updated (Line 6) if $\mathbf{X}^1_{t+1}$ is accepted: it increases (implying a higher penalization of unfeasible candidate solutions) when $\mathbf{X}^1_{t+1}$ is unfeasible and decreases otherwise. Our update of the Lagrange factor differs from the one in [3] in that it does not restrict λ_t to positive values. This modification appeared to be necessary for us to construct a homogeneous Markov chain whose stability implies linear convergence of the algorithm.

Similarly to the Lagrange factor, the penalty factor μ_t is updated when $\mathbf{X}^1_{t+1}$ is accepted (Line 7), where $\chi, k_1, k_2 \in \mathbb{R}^+_>$. The factor is increased when (i) the penalty term corresponding to $\mathbf{X}^1_{t+1}$ is smaller than the change in h value (first inequality in Line 7). This corresponds to the situation where the Lagrangian part, $f(\mathbf{x}) + \lambda g(\mathbf{x})$, appears to dominate $h(\mathbf{x})$. In this case we increase the penalization so that also the augmenting part, $\mu_t g^2(\mathbf{x})/2$, becomes visible to selection. The other situation where the penalty factor is increased is (ii) when the change in the distance to the constraint boundary $|\Delta g|$ (Line 4) is significantly smaller than the distance to the constraint boundary of the current solution $|g(\mathbf{X}_t)|$ (second inequality in Line 7). In this case, the penalization is increased to avoid premature stagnation when the search process is still far from the constraint boundary, as large values of μ_t guide the search more quickly towards $g(\mathbf{x}) = 0$. When conditions (i) and (ii) are not satisfied, μ_t is decreased to avoid an unnecessary ill-conditioning of the problem.

The updates of $\mathbf{X}_t$ and σ_t depend only on the ranking of h values of the candidate solutions. For λ_t and μ_t however, the algorithm explicitly uses h and g values of $\mathbf{X}_t$ and $\mathbf{X}_{t+1}$.

Algorithm 1 The $(1+1)$-ES with Augmented Lagrangian Constraint Handling

0 **given** $n \in \mathbb{N}_>$, $\chi, k_1, k_2 \in \mathbb{R}^+_>$
1 **initialize** $\mathbf{X}_0 \in \mathbb{R}^n$, $\sigma_0 \in \mathbb{R}^+_>$, $\lambda_0 \in \mathbb{R}$, $\mu_0 \in \mathbb{R}^+_>$, $t = 0$
2 **while** not happy
3 Compute $\mathbf{X}^1_{t+1} = \mathbf{X}_t + \sigma_t \mathbf{Z}_{t+1}$, where $\mathbf{Z}_{t+1} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{n\times n})$
4 Compute $\Delta g = g(\mathbf{X}^1_{t+1}) - g(\mathbf{X}_t)$
and $\Delta h = h(\mathbf{X}^1_{t+1}, \lambda_t, \mu_t) - h(\mathbf{X}_t, \lambda_t, \mu_t)$
5 **if** $\Delta h \leq 0$ **then**
6 $\lambda_{t+1} = \lambda_t + \mu_t g(\mathbf{X}^1_{t+1})$
7 $\mu_{t+1} = \begin{cases} \mu_t \chi^{1/4} & \text{if } \mu_t g^2(\mathbf{X}^1_{t+1}) < k_1 \frac{|\Delta h|}{n} \\ & \text{or } k_2 |\Delta g| < |g(\mathbf{X}_t)| \\ \mu_t \chi^{-1} & \text{otherwise} \end{cases}$
8 $\mathbf{X}_{t+1} = \mathbf{X}^1_{t+1}$
9 $\sigma_{t+1} = \sigma_t 2^{1/n}$
10 **else**
11 $\sigma_{t+1} = \sigma_t 2^{-1/(4n)}$
12 $t = t + 1$

Referring to (6), the transition function $\mathcal{F}^{(f,g)}$ of Algorithm 1 can be expressed as

$$\begin{aligned}\mathcal{F}^{(f,g)}((\mathbf{X}_t, \sigma_t, \lambda_t, \mu_t), \mathbf{U}_{t+1}) = \big(&\mathcal{G}_1((\mathbf{X}_t, \sigma_t), \varsigma * \mathbf{U}_{t+1}), \\ &\mathcal{G}_2(\sigma_t, \varsigma * \mathbf{U}_{t+1}), \mathcal{G}_3^{(f,g)}(\lambda_t, \mu_t, \mathbf{X}_t, \mathcal{G}_1((\mathbf{X}_t, \sigma_t), \varsigma * \mathbf{U}_{t+1})), \\ &\mathcal{G}_4^{(f,g)}(\mu_t, \lambda_t, \mathbf{X}_t, \mathcal{G}_1((\mathbf{X}_t, \sigma_t), \varsigma * \mathbf{U}_{t+1}))\big) \ ,\end{aligned} \tag{7}$$

where $\mathbf{U}_{t+1} = (\mathbf{Z}_{t+1}, \mathbf{0}) \in \mathbb{R}^{n\times 2}$ and

$$\varsigma = Ord(h(\mathbf{X}_t + \sigma_t [\mathbf{U}_{t+1}]_i, \lambda_t, \mu_t)_{i=1,2}) \tag{8}$$

is the permutation of indices of candidate solutions ordered according to h. Where relevant, we will explicitly write the dependence of ς on the variables used to compute candidate solutions and the fitness used to rank them (here, this would read $\varsigma^{h(\mathbf{x},\lambda_t,\mu_t)}_{(\mathbf{X}_t,\sigma_t)}$). The operator $*$ applies the permutation ς to $\mathbf{U}_{t+1}$ and returns the ranked vector $\varsigma * \mathbf{U}_{t+1} = ([\mathbf{U}_{t+1}]_{[\varsigma]_1}, [\mathbf{U}_{t+1}]_{[\varsigma]_2})$. Functions $\mathcal{G}_1$, $\mathcal{G}_2$, $\mathcal{G}_3$, and $\mathcal{G}_4$ compute the new state variables of the algorithm by updating the current state variables $\mathbf{X}_t$, σ_t, λ_t, and μ_t respectively. They are given by

$$\mathbf{X}_{t+1} = \mathcal{G}_1((\mathbf{X}_t, \sigma_t), \varsigma * \mathbf{U}_{t+1}) = \mathbf{X}_t + \sigma_t [\varsigma * \mathbf{U}_{t+1}]_1 \ , \tag{9}$$

$$\sigma_{t+1} = \mathcal{G}_2(\sigma_t, \varsigma * \mathbf{U}_{t+1}) = \sigma_t \underbrace{2^{-\frac{1}{4n} + \frac{5}{4n}\mathbf{1}_{\{[\varsigma * \mathbf{U}_{t+1}]_1 \neq \mathbf{0}\}}}}_{\eta^*(\varsigma * \mathbf{U}_{t+1})} \ , \tag{10}$$

where $\eta^*(\varsigma * \mathbf{U}_{t+1})$ is the step-size change (we will sometimes omit

the dependence on $\varsigma * \mathbf{U}_{t+1}$ for the sake of simplicity),

$$\lambda_{t+1} = \mathcal{G}_3^{(f,g)}(\lambda_t, \mu_t, \mathbf{X}_t, \mathbf{X}_{t+1}) = \lambda_t + \mu_t g(\mathbf{X}_{t+1}) \times \mathbf{1}_{\{[\varsigma * \mathbf{U}_{t+1}]_1 \neq \mathbf{0}\}} \ , \quad (11)$$

$$\mu_{t+1} = \mathcal{G}_4^{(f,g)}(\mu_t, \lambda_t, \mathbf{X}_t, \mathbf{X}_{t+1}) = \begin{cases} \mu_t \beta_t & \text{if } [\varsigma * \mathbf{U}_{t+1}]_1 \neq \mathbf{0} \\ \mu_t & \text{otherwise} \end{cases} \quad (12)$$

with

$$\beta_t = \begin{cases} \chi^{1/4} & \text{if } \mu_t g^2(\mathbf{X}_{t+1}) < k_1 \frac{|h(\mathbf{X}_{t+1}, \lambda_t, \mu_t) - h(\mathbf{X}_t, \lambda_t, \mu_t)|}{n} \\ & \text{or } k_2 |g(\mathbf{X}_{t+1}) - g(\mathbf{X}_t)| < |g(\mathbf{X}_t)| \\ \chi^{-1} & \text{otherwise} \ . \end{cases} \quad (13)$$

4.1 Invariance

We discuss here invariance with respect to transformations of the search space. We distinguish translation-invariance and scale-invariance.

Before giving the formal definitions of translation and scale-invariance, we remind the definition of a group homomorphism.

Definition 1. Let $(G, .)$ and $(H, *)$ be two groups. A function $\Phi : G \to H$ is a group homomorphism if for all $x, y \in G$, $\Phi(x.y) = \Phi(x) * \Phi(y)$.

Let $\mathcal{S}(\Omega)$ be the set of all bijective transformations from the state space Ω to itself and let $\text{Homo}((\mathbb{R}^n, +), (\mathcal{S}(\Omega), \circ))$ (respectively $\text{Homo}((\mathbb{R}^+_>, .), (\mathcal{S}(\Omega), \circ))$) be the set of group homomorphisms from $(\mathbb{R}^n, +)$ (respectively from $(\mathbb{R}^+_>, .)$) to $(\mathcal{S}(\Omega), \circ)$.

Definition 2. A randomized adaptive algorithm with transition function $\mathcal{F}^{(f,g)}$, where f is the objective function being minimized and g is the constraint function, is translation-invariant if there exists a group homomorphism $\Phi \in \text{Homo}((\mathbb{R}^n, +), (\mathcal{S}(\Omega), \circ))$ such that for any objective function f, for any constraint g, for any $\mathbf{x}_0 \in \mathbb{R}^n$, for any state $\mathbf{s} \in \Omega$, and for any $\boldsymbol{u} \in \mathbb{U}^p$,

$$\mathcal{F}^{(f(\mathbf{x}),g(\mathbf{x}))}(\mathbf{s}, \boldsymbol{u}) = \Phi(-\mathbf{x}_0) \left(\mathcal{F}^{(f(\mathbf{x}-\mathbf{x}_0),g(\mathbf{x}-\mathbf{x}_0))}(\Phi(\mathbf{x}_0)(\mathbf{s}), \boldsymbol{u}) \right) \ .$$

Informally, the previous definition means that if we transform the current state $\mathbf{s}_t$ of the algorithm via $\Phi(\mathbf{x}_0)$, perform one iteration to optimize $f(\mathbf{x} - \mathbf{x}_0)$ subject to $g(\mathbf{x} - \mathbf{x}_0) \leq 0$, and apply the inverse transformation $\Phi(-\mathbf{x}_0)$ to the resulting state, then we will recover the same state $\mathbf{s}_{t+1}$ as when starting from $\mathbf{s}_t$ and performing one iteration of the algorithm to optimize $f(\mathbf{x})$ subject to $g(\mathbf{x})$.

Definition 3. A randomized adaptive algorithm with transition function $\mathcal{F}^{(f,g)}$, where f is the objective function being minimized and g is the constraint, is scale-invariant if there exists a group homomorphism $\Phi \in \text{Homo}((\mathbb{R}^+_>, .), (\mathcal{S}(\Omega), \circ))$ such that for any objective function f, for any constraint g, for any $\alpha > 0$, for any state $\mathbf{s} \in \Omega$, and for any $\boldsymbol{u} \in \mathbb{U}^p$,

$$\mathcal{F}^{(f(\mathbf{x}),g(\mathbf{x}))}(\mathbf{s}, \boldsymbol{u}) = \Phi(1/\alpha) \left(\mathcal{F}^{(f(\alpha\mathbf{x}),g(\alpha\mathbf{x}))}(\Phi(\alpha)(\mathbf{s}), \boldsymbol{u}) \right) \ .$$

In the sequel, we prove that Algorithm 1 is translation and scale-invariant.

PROPOSITION 1. *Algorithm 1 is translation-invariant and the associated group homomorphism Φ is defined as*

$$\Phi(\boldsymbol{x}_0)(\boldsymbol{x}, \sigma, \lambda, \mu) = (\boldsymbol{x} + \boldsymbol{x}_0, \sigma, \lambda, \mu) \ , \quad (14)$$

for all $\boldsymbol{x}_0, \boldsymbol{x} \in \mathbb{R}^n$ and for all $\sigma, \lambda, \mu \in \mathbb{R}$.

PROOF. Consider the homomorphism defined in (14) and let $\mathbf{s}_t = (\mathbf{X}_t, \sigma_t, \lambda_t, \mu_t)$ and $\Phi(\mathbf{x}_0)(\mathbf{s}_t) = (\mathbf{X}'_t, \sigma'_t, \lambda'_t, \mu'_t)$. We have

$$h(\mathbf{X}_t + \sigma_t [\mathbf{U}_{t+1}]_i, \lambda_t, \mu_t) = h(\mathbf{X}'_t + \sigma'_t [\mathbf{U}_{t+1}]_i - \mathbf{x}_0, \lambda'_t, \mu'_t) \ ,$$

where $\mathbf{U}_{t+1} = (\mathbf{Z}_{t+1}, \mathbf{0})$. Consequently, the same permutation ς is obtained when ranking candidate solutions $\mathbf{X}'_t + \sigma'_t [\mathbf{U}_{t+1}]_i$, $i = 1, 2$, on $h(\mathbf{x} - \mathbf{x}_0, \lambda, \mu)$ than when ranking candidate solutions $\mathbf{X}_t + \sigma_t [\mathbf{U}_{t+1}]_i$, $i = 1, 2$, on $h(\mathbf{x}, \lambda, \mu)$. Therefore, according to (7), $\mathcal{F}^{(f(\mathbf{x}-\mathbf{x}_0),g(\mathbf{x}-\mathbf{x}_0))}(\Phi(\mathbf{x}_0)(\mathbf{s}_t), \mathbf{U}_{t+1})$ writes

$$\mathbf{X}'_{t+1} = \mathcal{G}_1((\mathbf{X}'_t, \sigma'_t), \varsigma * \mathbf{U}_{t+1}) = \mathcal{G}_1((\mathbf{X}_t, \sigma_t), \varsigma * \mathbf{U}_{t+1}) + \mathbf{x}_0 \ , \quad (15)$$

$$\sigma'_{t+1} = \mathcal{G}_2(\sigma'_t, \varsigma * \mathbf{U}_{t+1}) = \mathcal{G}_2(\sigma_t, \varsigma * \mathbf{U}_{t+1}) \ ,$$

$$\lambda'_{t+1} = \mathcal{G}_3^{(f(\mathbf{x}-\mathbf{x}_0),g(\mathbf{x}-\mathbf{x}_0))}(\lambda'_t, \mu'_t, \mathbf{X}'_t, \mathbf{X}'_{t+1}) = \mathcal{G}_3^{(f(\mathbf{x}),g(\mathbf{x}))}(\lambda_t, \mu_t, \mathbf{X}_t, \mathbf{X}_{t+1}) \ ,$$

$$\mu'_{t+1} = \mathcal{G}_4^{(f(\mathbf{x}-\mathbf{x}_0),g(\mathbf{x}-\mathbf{x}_0))}(\mu'_t, \lambda'_t, \mathbf{X}'_t, \mathbf{X}'_{t+1}) = \mathcal{G}_4^{(f(\mathbf{x}),g(\mathbf{x}))}(\mu_t, \lambda_t, \mathbf{X}_t, \mathbf{X}_{t+1}) \ .$$

We recover $\mathcal{F}^{(f(\mathbf{x}),g(\mathbf{x}))}((\mathbf{X}_t, \sigma_t, \lambda_t, \mu_t), \mathbf{U}_{t+1})$ by applying the inverse transformation $\Phi(-\mathbf{x}_0)$ to $(\mathbf{X}'_{t+1}, \sigma'_{t+1}, \lambda'_{t+1}, \mu'_{t+1})$. □

PROPOSITION 2. *Algorithm 1 is scale-invariant and the associated group homomorphism Φ is defined as*

$$\Phi(\alpha)(\boldsymbol{x}, \sigma, \lambda, \mu) = (\boldsymbol{x}/\alpha, \sigma/\alpha, \lambda, \mu) \ , \quad (16)$$

for all $\alpha \in \mathbb{R}^+_>$, for all $\boldsymbol{x} \in \mathbb{R}^n$, and for all $\sigma, \lambda, \mu \in \mathbb{R}$.

PROOF. Let $\mathbf{s}_t = (\mathbf{X}_t, \sigma_t, \lambda_t, \mu_t)$ and $\Phi(\alpha)(\mathbf{s}_t) = (\mathbf{X}'_t, \sigma'_t, \lambda'_t, \mu'_t)$. We use the same idea as in the previous proof to show that the same permutation ς is obtained when ranking candidate solutions $\mathbf{X}'_t + \sigma'_t [\mathbf{U}_{t+1}]_i$, $i = 1, 2$, on $h(\alpha\mathbf{x}, \lambda, \mu)$ than when ranking candidate solutions $\mathbf{X}_t + \sigma_t [\mathbf{U}_{t+1}]_i$, $i = 1, 2$, on $h(\mathbf{x}, \lambda, \mu)$. Therefore, according to (7), $\mathcal{F}^{(f(\alpha\mathbf{x}),g(\alpha\mathbf{x}))}(\Phi(\mathbf{s}_t), \mathbf{U}_{t+1})$ writes

$$\mathbf{X}'_{t+1} = \mathcal{G}_1((\mathbf{X}'_t, \sigma'_t), \varsigma * \mathbf{U}_{t+1}) = \frac{1}{\alpha} \mathcal{G}_1((\mathbf{X}_t, \sigma_t), \varsigma * \mathbf{U}_{t+1}) \ , \quad (17)$$

$$\sigma'_{t+1} = \mathcal{G}_2(\sigma'_t, \varsigma * \mathbf{U}_{t+1}) = \frac{1}{\alpha} \mathcal{G}_2(\sigma_t, \varsigma * \mathbf{U}_{t+1}) \ , \quad (18)$$

$$\lambda'_{t+1} = \mathcal{G}_3^{(f(\alpha\mathbf{x}),g(\alpha\mathbf{x}))}(\lambda'_t, \mu'_t, \mathbf{X}'_t, \mathbf{X}'_{t+1}) = \mathcal{G}_3^{(f(\mathbf{x}),g(\mathbf{x}))}(\lambda_t, \mu_t, \mathbf{X}_t, \mathbf{X}_{t+1}) \ ,$$

$$\mu'_{t+1} = \mathcal{G}_4^{(f(\alpha\mathbf{x}),g(\alpha\mathbf{x}))}(\mu'_t, \lambda'_t, \mathbf{X}'_t, \mathbf{X}'_{t+1}) = \mathcal{G}_4^{(f(\mathbf{x}),g(\mathbf{x}))}(\mu_t, \lambda_t, \mathbf{X}_t, \mathbf{X}_{t+1}) \ .$$

We recover $\mathcal{F}^{(f(\mathbf{x}),g(\mathbf{x}))}(\mathbf{s}_t, \mathbf{U}_{t+1})$ by applying the inverse transformation $\Phi(1/\alpha)$ to $(\mathbf{X}'_{t+1}, \sigma'_{t+1}, \lambda'_{t+1}, \mu'_{t+1})$. □

5. ANALYSIS

In this section, we investigate the behavior of Algorithm 1 on the augmented Lagrangian h. We start by showing that given a particular condition is satisfied by h, we can construct a homogeneous Markov chain from the state variables of the algorithm, by exploiting its invariance properties as well as the updates of λ_t and μ_t. In the second part, we illustrate how the stability of the constructed Markov chain results in linear convergence of $\mathbf{X}_t$ towards the optimum $\mathbf{x}_{\text{opt}}$, as well as linear convergence of λ_t and σ_t towards λ_{opt} and 0 respectively.

5.1 Homogeneous Markov Chain

Before presenting the Markov chain, we extend the definition of positive homogeneity with respect to zero to any vector $\mathbf{x}^*$.

Definition 4. A function $p : X \to Y$ is positive homogeneous of degree $k > 0$ with respect to $\mathbf{x}^* \in X$ if for all $\alpha > 0$ and for all $\mathbf{x} \in X$,

$$p(\mathbf{x}^* + \alpha\mathbf{x}) = \alpha^k p(\mathbf{x}^* + \mathbf{x}) \ . \tag{19}$$

By taking $\mathbf{x}^* = 0$, we recover the standard definition of positive homogeneity.

Our linear constraint function $g(\mathbf{x}) = \mathrm{b}^T\mathbf{x} + c$ is positive homogeneous of degree 1 with respect to any $\mathbf{x}^* \in \mathbb{R}^n$ such that $g(\mathbf{x}^*) = 0$. The sphere function $p_{\text{sphere}}(\mathbf{x}) = (\mathbf{x} - \mathbf{x}^*)^T(\mathbf{x} - \mathbf{x}^*)$ is also positive homogeneous of degree 2 with respect to $\mathbf{x}^*$.

We will now define two random variables, $\mathbf{Y}_t$ and Λ_t, and prove that if the augmented Lagrangian h satisfies the condition stated below in (21), then $(\mathbf{Y}_t, \Lambda_t, \mu_t)_{t\in\mathbb{N}}$ is a Markov chain. For the proof, we use transition and scale-invariance along with the updates of λ_t and μ_t.

PROPOSITION 3. *Consider the $(1+1)$-ES with augmented Lagrangian constraint handling optimizing the augmented Lagrangian h defined in (5). Let $(\boldsymbol{X}_t, \sigma_t, \lambda_t, \mu_t)_{t\in\mathbb{N}}$ be the Markov chain associated to this ES and let $(\boldsymbol{U}_t)_{t\in\mathbb{N}}$ be the sequence of i.i.d. random vectors where $\boldsymbol{U}_{t+1} = (\boldsymbol{Z}_{t+1}, \mathbf{0}) \in \mathbb{R}^{n\times 2}$ and $\boldsymbol{Z}_{t+1} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{n\times n})$. Let*

$$\boldsymbol{Y}_t = \frac{\boldsymbol{X}_t - \bar{\boldsymbol{x}}}{\sigma_t} \ \text{ and } \ \Lambda_t = \frac{\lambda_t - \bar{\lambda}}{\sigma_t} \ . \tag{20}$$

Then, if the function $\mathcal{D}h_{\bar{\mathbf{x}},\bar{\lambda},\mu} : \mathbb{R}^{n+1} \to \mathbb{R}$ defined as follows

$$\mathcal{D}h_{\bar{\mathbf{x}},\bar{\lambda},\mu}(\boldsymbol{x}, \lambda) = h(\boldsymbol{x}, \lambda, \mu) - h(\bar{\boldsymbol{x}}, \bar{\lambda}, \mu) \ , \tag{21}$$

where $\boldsymbol{x}, \bar{\boldsymbol{x}} \in \mathbb{R}^n$, $\lambda, \bar{\lambda} \in \mathbb{R}$, and $g(\bar{\boldsymbol{x}}) = 0$, is positive homogeneous of degree 2 with respect to $(\bar{\boldsymbol{x}}, \bar{\lambda})$, then $(\boldsymbol{Y}_t, \Lambda_t, \mu_t)_{t\in\mathbb{N}}$ is a homogeneous Markov chain defined independently of $(\boldsymbol{X}_t, \sigma_t, \lambda_t, \mu_t)$ as

$$\boldsymbol{Y}_{t+1} = \mathcal{G}_1((\boldsymbol{Y}_t, 1), \varsigma * \boldsymbol{U}_{t+1})/\eta^* \ , \tag{22}$$

$$\Lambda_{t+1} = (\mathcal{G}_3^{(f(\boldsymbol{x}+\bar{\boldsymbol{x}}),g(\boldsymbol{x}+\bar{\boldsymbol{x}}))}(\Lambda_t + \bar{\lambda}, \mu_t, \boldsymbol{Y}_t, \eta^*\boldsymbol{Y}_{t+1}) - \bar{\lambda})/\eta^* \ , \tag{23}$$

$$\mu_{t+1} = \mathcal{G}_4^{(f(\boldsymbol{x}+\bar{\boldsymbol{x}}),g(\boldsymbol{x}+\bar{\boldsymbol{x}}))}(\mu_t, \Lambda_t + \bar{\lambda}, \boldsymbol{Y}_t, \eta^*\boldsymbol{Y}_{t+1}) \ , \tag{24}$$

where $\eta^* = \eta^*(\varsigma * \boldsymbol{U}_{t+1})$ and

$$\varsigma = \mathit{Ord}(h(\boldsymbol{Y}_t + [\boldsymbol{U}_{t+1}]_i + \bar{\boldsymbol{x}}, \Lambda_t + \bar{\lambda}, \mu_t)_{i=1,2}) \ . \tag{25}$$

PROOF. We show that $\mathbf{Y}_{t+1}$, Λ_{t+1}, and μ_{t+1} only depend on $\mathbf{Y}_t$, Λ_t, μ_t and i.i.d. random variables $\mathbf{U}_{t+1}$, and therefore that $(\mathbf{Y}_t, \Lambda_t, \mu_t)_{t\in\mathbb{N}}$ is a homogeneous Markov chain.

Given the definitions of $\mathbf{Y}_t$ and Λ_t in Proposition 3, we can write

$$h(\mathbf{X}_t + \sigma_t[\mathbf{U}_{t+1}]_i, \lambda_t, \mu_t) = h(\sigma_t(\mathbf{Y}_t + [\mathbf{U}_{t+1}]_i) + \bar{\mathbf{x}}, \sigma_t\Lambda_t + \bar{\lambda}, \mu_t) \ .$$

Consider ranking the elements of the set

$$\{h(\sigma_t(\mathbf{Y}_t + [\mathbf{U}_{t+1}]_i) + \bar{\mathbf{x}}, \sigma_t\Lambda_t + \bar{\lambda}, \mu_t)\}_{i=1,2} \ .$$

We obtain the same permutation when ranking the elements of

$$\{\mathcal{D}h_{\bar{\mathbf{x}},\bar{\lambda},\mu_t}(\sigma_t(\mathbf{Y}_t + [\mathbf{U}_{t+1}]_i) + \bar{\mathbf{x}}, \sigma_t\Lambda_t + \bar{\lambda})\}_{i=1,2} \ ,$$

where $\mathcal{D}h_{\bar{\mathbf{x}},\bar{\lambda},\mu_t}$ is defined in (21). $\mathcal{D}h_{\bar{\mathbf{x}},\bar{\lambda},\mu_t}$ being positive homogeneous with respect to $(\bar{\mathbf{x}}, \bar{\lambda})$, the ranking is the same on

$$\{\mathcal{D}h_{\bar{\mathbf{x}},\bar{\lambda},\mu_t}(\mathbf{Y}_t + [\mathbf{U}_{t+1}]_i + \bar{\mathbf{x}}, \Lambda_t + \bar{\lambda})\}_{i=1,2}$$

and, consequently, on

$$\{h(\mathbf{Y}_t + [\mathbf{U}_{t+1}]_i + \bar{\mathbf{x}}, \Lambda_t + \bar{\lambda})\}_{i=1,2} \ .$$

Therefore, the same permutation ς defined in (25) is obtained when ranking the candidate solutions $\mathbf{X}_t + \sigma_t[\mathbf{U}_{t+1}]_i$, $i = 1, 2$, on $h(\mathbf{x}, \lambda_t, \mu_t)$ than when ranking the candidate solutions $\mathbf{Y}_t + [\mathbf{U}_{t+1}]_i$ on $h(\mathbf{x} + \bar{\mathbf{x}}, \Lambda_t + \bar{\lambda}, \mu_t)$. It follows that

$$\begin{aligned}\mathbf{Y}_{t+1} = \frac{\mathbf{X}_{t+1} - \bar{\mathbf{x}}}{\sigma_{t+1}} &= \frac{\mathcal{G}_1((\mathbf{X}_t, \sigma_t), \varsigma * \mathbf{U}_{t+1}) - \bar{\mathbf{x}}}{\mathcal{G}_2(\sigma_t, \varsigma * \mathbf{U}_{t+1})} \\ &= \mathcal{G}_1((\mathbf{Y}_t, 1), \varsigma * \mathbf{U}_{t+1})/\eta^* \ ,\end{aligned} \tag{26}$$

where we used scale-invariance properties of $\mathcal{G}_1$ and $\mathcal{G}_2$ ((17) and (18)) and translation-invariance property of $\mathcal{G}_1$ in (15).

On the other hand, we have

$$\Lambda_{t+1} = \frac{\lambda_{t+1} - \bar{\lambda}}{\sigma_{t+1}} = \frac{\mathcal{G}_3^{(f(\mathbf{x}),g(\mathbf{x}))}(\lambda_t, \mu_t, \mathbf{X}_t, \mathbf{X}_{t+1}) - \bar{\lambda}}{\mathcal{G}_2(\sigma_t, \varsigma * \mathbf{U}_{t+1})} \ , \tag{27}$$

where $\mathcal{G}_3$ is given in (11). Using scale-invariance of $\mathcal{G}_2$ and positive homogeneity of g with respect to $\bar{\mathbf{x}}$, it follows that

$$\begin{aligned}\mathcal{G}_3^{(f(\mathbf{x}),g(\mathbf{x}))}(\lambda_t, \mu_t, \mathbf{X}_t, \mathbf{X}_{t+1}) - \bar{\lambda} = \sigma_t\big(\mathcal{G}_3^{(f(\mathbf{x}+\bar{\mathbf{x}}),g(\mathbf{x}+\bar{\mathbf{x}}))}(\Lambda_t + \bar{\lambda}, \\ \mu_t, \mathbf{Y}_t, \eta^*\mathbf{Y}_{t+1}) - \bar{\lambda}\big) \ .\end{aligned}$$

Replacing in (27), we obtain (23).

Remark 2. With the update of λ_t used in [3] ($\lambda_{t+1} = \max(0, \lambda_t + \mu_t g(\mathbf{X}_t + \sigma_t\mathbf{Z}_{t+1})$ if $\Delta h \leq 0$, λ_t otherwise), Λ_{t+1} cannot be written as a function of $(\mathbf{Y}_t, \Lambda_t, \mu_t)$. Indeed, because of the max function, one cannot get rid of σ_t.

μ_{t+1} is given in (24). $\mathcal{D}h_{\bar{\mathbf{x}},\bar{\lambda},\mu_t}$ is positive homogeneous of degree 2 with respect to $(\bar{\mathbf{x}}, \bar{\lambda})$. Therefore, according to Definition 4 and for $\alpha = \sigma_t$,

$$\mathcal{D}h_{\bar{\mathbf{x}},\bar{\lambda},\mu_t}(\mathbf{X}_{t+1}, \lambda_t) = \sigma_t^2\mathcal{D}h_{\bar{\mathbf{x}},\bar{\lambda},\mu_t}(\eta^*\mathbf{Y}_{t+1} + \bar{\mathbf{x}}, \Lambda_t + \bar{\lambda}) \tag{28}$$

and

$$\mathcal{D}h_{\bar{\mathbf{x}},\bar{\lambda},\mu_t}(\mathbf{X}_t, \lambda_t) = \sigma_t^2\mathcal{D}h_{\bar{\mathbf{x}},\bar{\lambda},\mu_t}(\mathbf{Y}_t + \bar{\mathbf{x}}, \Lambda_t + \bar{\lambda}) \ , \tag{29}$$

where we used (20). Subtracting (29) from (28), we get

$$\begin{aligned}h(\mathbf{X}_{t+1}, \lambda_t, \mu_t) - h(\mathbf{X}_t, \lambda_t, \mu_t) = \sigma_t^2\big(h(\eta^*\mathbf{Y}_{t+1} + \bar{\mathbf{x}}, \Lambda_t + \bar{\lambda}, \mu_t) \\ - h(\mathbf{Y}_t + \bar{\mathbf{x}}, \Lambda_t + \bar{\lambda}, \mu_t)\big) \ .\end{aligned} \tag{30}$$

Using (30) and positive homogeneity of g with respect to $\bar{\mathbf{x}}$, we get

$$\beta_t = \begin{cases} \chi^{1/4} & \text{if } \mu_t g^2(\eta^*\mathbf{Y}_{t+1} + \bar{\mathbf{x}}) < k_1 \\ & \quad \times \frac{|h(\eta^*\mathbf{Y}_{t+1} + \bar{\mathbf{x}}, \Lambda_t + \bar{\lambda}, \mu_t) - h(\mathbf{Y}_t + \bar{\mathbf{x}}, \Lambda_t + \bar{\lambda}, \mu_t)|}{n} \\ & \text{or } k_2|g(\eta^*\mathbf{Y}_{t+1} + \bar{\mathbf{x}}) - g(\mathbf{Y}_t + \bar{\mathbf{x}})| < |g(\mathbf{Y}_t + \bar{\mathbf{x}})| \\ \chi^{-1} & \text{otherwise} \ , \end{cases}$$

therefore, (24) follows. □

The result in Proposition 3 is particularly interesting if $\bar{\mathbf{x}}$ and $\bar{\lambda}$ correspond to the optimum of the constrained problem, $\mathbf{x}_{\text{opt}}$, and to the Lagrange multiplier, λ_{opt}, respectively. In this case, one can express the convergence rate of the algorithm towards $\mathbf{x}_{\text{opt}}$ as a function of the homogeneous Markov chain $(\mathbf{Y}_t, \Lambda_t, \mu_t)_{t\in\mathbb{N}}$, where $\mathbf{Y}_t = \frac{\mathbf{X}_t - \mathbf{x}_{\text{opt}}}{\sigma_t}$ and $\Lambda_t = \frac{\lambda_t - \lambda_{\text{opt}}}{\sigma_t}$. The LLN can then be applied to prove linear convergence if the Markov chain satisfies some stability conditions, which are further discussed in Section 5.

COROLLARY 1. *Let $(\boldsymbol{X}_t, \sigma_t, \lambda_t, \mu_t)_{t\in\mathbb{N}}$ be the Markov chain associated to Algorithm 1 optimizing the augmented Lagrangian h in (5), where f is a convex quadratic function defined as*

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{H}\boldsymbol{x} \ , \tag{31}$$

with $\boldsymbol{H} \in \mathbb{R}^{n\times n}$ a symmetric positive-definite matrix. Let $\boldsymbol{Y}_t = \frac{\boldsymbol{X}_t - \boldsymbol{x}_{opt}}{\sigma_t}$ and $\Lambda_t = \frac{\lambda_t - \lambda_{\text{opt}}}{\sigma_t}$, where $\boldsymbol{x}_{opt}$ is the optimum and λ_{opt} is the associated Lagrange multiplier. Then $(\boldsymbol{Y}_t, \Lambda_t, \mu_t)_{t\in\mathbb{N}}$ is a homogeneous Markov chain defined independently of $(\boldsymbol{X}_t, \sigma_t, \lambda_t, \mu_t)$ as in Equations 22, 23, 24, and 25, where $\bar{\boldsymbol{x}} = \boldsymbol{x}_{opt}$ and $\bar{\lambda} = \lambda_{\text{opt}}$.

Before moving to the proof, we remind that for f convex quadratic and g linear, KKT conditions are also sufficient conditions of optimality, that is, a point satisfying KKT conditions is also an optimum of the constrained problem (see Theorem 16.4 in [11]). Since the problem is unimodal, KKT conditions are satisfied only for $\mathbf{x}_{\text{opt}}$ and λ_{opt}, and we have

$$\nabla_{\mathbf{x}} f(\mathbf{x}_{\text{opt}}) + \lambda_{\text{opt}} \nabla_{\mathbf{x}} g(\mathbf{x}_{\text{opt}}) = \mathbf{0} \ . \tag{32}$$

PROOF. We show that for $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x}$, $\mathcal{D}h_{\mathbf{x}_{\text{opt}},\lambda_{\text{opt}},\mu}$ in (21) is positive homogeneous of degree 2 with respect to $(\mathbf{x}_{\text{opt}}, \lambda_{\text{opt}}) \in \mathbb{R}^{n+1}$ and therefore, by virtue of Proposition 3, $(\mathbf{Y}_t, \Lambda_t, \mu_t)_{t\in\mathbb{N}}$ is a homogeneous Markov chain. We have by definition of h

$$h(\mathbf{x}_{\text{opt}} + \alpha\mathbf{x}, \lambda_{\text{opt}} + \alpha\lambda, \mu) = \underbrace{f(\mathbf{x}_{\text{opt}} + \alpha\mathbf{x})}_{A} + \underbrace{(\lambda_{\text{opt}} + \alpha\lambda) g(\mathbf{x}_{\text{opt}} + \alpha\mathbf{x})}_{B} + \underbrace{\frac{\mu}{2} g^2(\mathbf{x}_{\text{opt}} + \alpha\mathbf{x})}_{C} \ .$$

Given $\nabla_{\mathbf{x}} f(\mathbf{y}) = \mathbf{y}^T\mathbf{H}$ and $\nabla_{\mathbf{x}} g(\mathbf{y}) = \mathrm{b}^T$, it follows that

$$A = \alpha^2 f(\mathbf{x}_{\text{opt}} + \mathbf{x}) + (1-\alpha^2) f(\mathbf{x}_{\text{opt}}) + \alpha(1-\alpha)\nabla_{\mathbf{x}} f(\mathbf{x}_{\text{opt}})\mathbf{x} \ ,$$

$$B = \alpha^2(\lambda_{\text{opt}} + \lambda) g(\mathbf{x}_{\text{opt}} + \mathbf{x}) + \alpha(1-\alpha)\lambda_{\text{opt}} \nabla_{\mathbf{x}} g(\mathbf{x}_{\text{opt}})\mathbf{x} \ ,$$

$$C = \alpha^2 \frac{\mu}{2} g^2(\mathbf{x}_{\text{opt}} + \mathbf{x}) \ .$$

Therefore

$$h(\mathbf{x}_{\text{opt}} + \alpha\mathbf{x}, \lambda_{\text{opt}} + \alpha\lambda, \mu) = \alpha^2 h(\mathbf{x}_{\text{opt}} + \mathbf{x}, \lambda_{\text{opt}} + \lambda, \mu) + (1-\alpha^2) f(\mathbf{x}_{\text{opt}}) + \alpha(1-\alpha)(\nabla_{\mathbf{x}} f(\mathbf{x}_{\text{opt}}) + \lambda_{\text{opt}} \nabla_{\mathbf{x}} g(\mathbf{x}_{\text{opt}}))\mathbf{x} \ .$$

Using (32) and the fact that the constraint g is active at $\mathbf{x}_{\text{opt}}$, implying that $h(\mathbf{x}_{\text{opt}}, \lambda_{\text{opt}}, \mu) = f(\mathbf{x}_{\text{opt}})$, we get

$$\mathcal{D}h_{\mathbf{x}_{\text{opt}},\lambda_{\text{opt}},\mu}(\mathbf{x}_{\text{opt}} + \alpha\mathbf{x}, \lambda_{\text{opt}} + \alpha\lambda) = \alpha^2 \mathcal{D}h_{\mathbf{x}_{\text{opt}},\lambda_{\text{opt}},\mu}(\mathbf{x}_{\text{opt}} + \mathbf{x}, \lambda_{\text{opt}} + \lambda) \ .$$

□

Figure 2 shows contour lines of $\mathcal{D}h_{x_{\text{opt}},\lambda_{\text{opt}},\mu}(x, \lambda)$ (21), where the augmented Lagrangian h is defined for a particular convex quadratic function, the sphere $f(x) = \frac{1}{2}x^2$, $x \in \mathbb{R}$, and the constraint function $g(x) = -x + 1$. The penalty factor $\mu = 1$. In this setting, $x_{\text{opt}} = 1$ and $\lambda_{\text{opt}} = 1$. We can see from the figure that the function is scaling-invariant with respect to $(x_{\text{opt}}, \lambda_{\text{opt}})$: if we zoom in around the point $(x_{\text{opt}}, \lambda_{\text{opt}})$, we will still see the same contour lines. Algorithm 1 optimizes the function whose values correspond to a horizontal cut in the graph, that is, to a fixed value of λ. The intersection between the horizontal line $\lambda = \lambda_i$ and the blue line corresponds to $\min_x \mathcal{D}h_{x_{\text{opt}},\lambda_{\text{opt}},\mu}(x, \lambda_i, \mu)$ where $\arg\min_x \mathcal{D}h_{x_{\text{opt}},\lambda_{\text{opt}},\mu}(x, \lambda_i, \mu)$ can be read on the x-axis. For $\lambda = \lambda_{\text{opt}}$, the intersection happens in 0 and the corresponding value on the x-axis is $x_{\text{opt}} = 1$.

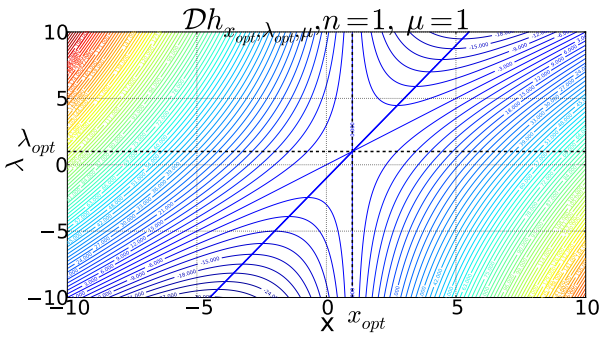

Figure 2: Contour lines of $\mathcal{D}h_{x_{\text{opt}},\lambda_{\text{opt}},\mu}$ for $f(x) = \frac{1}{2}x^2$, $g(\mathbf{x}) = -x + 1$, and $\mu = 1$. The vertical (respectively horizontal) dotted black line shows $x_{\text{opt}} = 1$ (respectively $\lambda_{\text{opt}} = 1$). Points where the solid blue line intersects the contour lines represent $\min_x \mathcal{D}h_{x_{\text{opt}},\lambda_{\text{opt}},\mu}(x, \lambda)$ for the corresponding λ.

5.2 Sufficient Conditions for Linear Convergence

Let us consider Algorithm 1 optimizing the augmented Lagrangian h from (5) such that the function $\mathcal{D}h_{\mathbf{x}_{\text{opt}},\lambda_{\text{opt}},\mu_t}$ defined in (21) is positive homogeneous of degree 2 with respect to $(\mathbf{x}_{\text{opt}}, \lambda_{\text{opt}})$, where $\mathbf{x}_{\text{opt}}$ is the optimum of the problem and λ_{opt} is the associated Lagrange multiplier. Let $(\mathbf{X}_t, \sigma_t, \lambda_t, \mu_t)_{t\in\mathbb{N}}$ be the Markov chain generated by the algorithm. Under these assumptions, let $(\mathbf{Y}_t, \Lambda_t, \mu_t)_{t\in\mathbb{N}}$ be the homogeneous Markov chain defined in Proposition 3. The log-progress $\ln \frac{\|\mathbf{X}_{t+1} - \mathbf{x}_{\text{opt}}\|}{\|\mathbf{X}_t - \mathbf{x}_{\text{opt}}\|}$ can be expressed as a function of $(\mathbf{Y}_t, \Lambda_t, \mu_t)$ as follows

$$\ln \frac{\|\mathbf{X}_{t+1} - \mathbf{x}_{\text{opt}}\|}{\|\mathbf{X}_t - \mathbf{x}_{\text{opt}}\|} = \ln \frac{\|\mathbf{Y}_{t+1}\|}{\|\mathbf{Y}_t\|} \eta^*(\varsigma * \mathbf{U}_{t+1}) \ , \tag{33}$$

where ς and η^* are defined in (25) and (10) respectively. By taking the sum then the limit of the average, we obtain the convergence rate

$$\lim_{t\to\infty} \frac{1}{t} \sum_{k=0}^{t-1} \ln \frac{\|\mathbf{X}_{k+1} - \mathbf{x}_{\text{opt}}\|}{\|\mathbf{X}_k - \mathbf{x}_{\text{opt}}\|} = \lim_{t\to\infty} \frac{1}{t} \sum_{k=0}^{t-1} \ln \frac{\|\mathbf{Y}_{k+1}\|}{\|\mathbf{Y}_k\|} \times \eta^*(\varsigma * \mathbf{U}_{k+1}) \ . \tag{34}$$

If the Markov chain $(\mathbf{Y}_t, \Lambda_t, \mu_t)_{t\in\mathbb{N}}$ is φ-irreducible and positive Harris-recurrent, then a LLN can be applied to the left-hand side of (34) to show almost sure linear convergence.

Before stating our theorem, we define $\mathcal{R}(\Phi)$, $\Phi = (\Phi_1, \Phi_2, \Phi_3)$, as the expectation of $\ln \eta^*(\varsigma^{h(\mathbf{x}+\mathbf{x}_{\text{opt}}, \Phi_2 + \lambda_{\text{opt}}, \Phi_3)}_{(\Phi_1,1)} * \mathbf{U})$ for $\mathbf{U} \sim p_{\mathbf{U}}$.

$$\mathcal{R}(\Phi) = \mathrm{E}\left(\ln \eta^*(\varsigma^{h(\mathbf{x}+\mathbf{x}_{\text{opt}}, \Phi_2 + \lambda_{\text{opt}}, \Phi_3)}_{(\Phi_1,1)} * \mathbf{U})\right) = \int \ln \eta^*(\varsigma^{h(\mathbf{x}+\mathbf{x}_{\text{opt}}, \Phi_2 + \lambda_{\text{opt}}, \Phi_3)}_{(\Phi_1,1)} * \boldsymbol{u})\, p_{\mathbf{U}}(\boldsymbol{u}) d\boldsymbol{u} \ . \tag{35}$$

We also recall Theorem 17.0.1 from [10], which gives sufficient conditions for the application of the LLN.

THEOREM 1 (THEOREM 17.0.1 FROM [10]). *Assume that $\boldsymbol{X}$ is a positive Harris-recurrent chain with invariant probability π. Then, the LLN holds for any q such that $\pi(|q|) = \int |q(\boldsymbol{x})| \pi(d\boldsymbol{x}) < \infty$, that is, for any initial state $\boldsymbol{X}_0$, $\lim_{t\to\infty} \frac{1}{t} \sum_{k=0}^{t-1} q(\boldsymbol{X}_k) = \pi(q)$ almost surely.*

THEOREM 2. *Let $(\boldsymbol{X}_t, \sigma_t, \lambda_t, \mu_t)_{t\in\mathbb{N}}$ be the Markov chain associated to Algorithm 1 optimizing the augmented Lagrangian h*

such that the function $\mathcal{D}h_{\boldsymbol{x}_{opt},\lambda_{\text{opt}},\mu_t}$ defined in (21) is positive homogeneous of degree 2 with respect to $(\boldsymbol{x}_{opt}, \lambda_{\text{opt}})$ (the optimum and the Lagrange multiplier respectively). Let $(\mathbf{Y}_t, \Lambda_t, \mu_t)_{t\in\mathbb{N}}$ be the Markov chain defined in Proposition 3 and assume that it is positive Harris-recurrent with invariant probability measure π, that $E_\pi(|\ln \|[\Phi]_1\||) < \infty$, $E_\pi(|(\ln |[\Phi]_2|)|) < \infty$, and $E_\pi(\mathcal{R}(\Phi)) < \infty$. Then, for all $\boldsymbol{X}_0$, for all σ_0, for all λ_0, and for all μ_0, linear convergence holds asymptotically almost surely, that is

$$\lim_{t\to\infty} \frac{1}{t} \ln \frac{\|\boldsymbol{X}_t - \boldsymbol{x}_{opt}\|}{\|\boldsymbol{X}_0 - \boldsymbol{x}_{opt}\|} = \lim_{t\to\infty} \frac{1}{t} \ln \frac{|\lambda_t - \lambda_{\text{opt}}|}{|\lambda_0 - \lambda_{\text{opt}}|} = \lim_{t\to\infty} \frac{1}{t} \ln \frac{\sigma_t}{\sigma_0} = -CR \ , \quad (36)$$

where

$$-CR = E_\pi(\mathcal{R}(\Phi)) = \int \mathcal{R}(\Phi)\pi(d\Phi) \ . \quad (37)$$

PROOF. Using the property of the logarithm, we have

$$\lim_{t\to\infty} \frac{1}{t} \ln \frac{\|\mathbf{X}_t - \mathbf{x}_{\text{opt}}\|}{\|\mathbf{X}_0 - \mathbf{x}_{\text{opt}}\|} = \lim_{t\to\infty} \frac{1}{t} \sum_{k=0}^{t-1} \ln \frac{\|\mathbf{X}_{k+1} - \mathbf{x}_{\text{opt}}\|}{\|\mathbf{X}_k - \mathbf{x}_{\text{opt}}\|} \ .$$

Then, using (34), we obtain

$$\lim_{t\to\infty} \frac{1}{t} \ln \frac{\|\mathbf{X}_t - \mathbf{x}_{\text{opt}}\|}{\|\mathbf{X}_0 - \mathbf{x}_{\text{opt}}\|} = \lim_{t\to\infty} \frac{1}{t} \sum_{k=0}^{t-1} \ln \|\mathbf{Y}_{k+1}\| - \lim_{t\to\infty} \frac{1}{t} \sum_{k=0}^{t-1} \ln \|\mathbf{Y}_k\| + \lim_{t\to\infty} \frac{1}{t} \sum_{k=0}^{t-1} \ln \eta^*(\varsigma * \mathbf{U}_{k+1}) \ . \quad (38)$$

Since $(\mathbf{Y}_t, \Lambda_t, \mu_t)_{t\in\mathbb{N}}$ is positive Harris-recurrent with an invariant probability measure π, it is possible to apply the LLN to the right-hand side of (38). Knowing that $\varsigma = \varsigma_{(\mathbf{Y}_t,1)}^{h(\mathbf{x}+\mathbf{x}_{\text{opt}},\Lambda_t+\lambda_{\text{opt}},\mu_t)}$, it follows

$$\lim_{t\to\infty} \frac{1}{t} \ln \frac{\|\mathbf{X}_t - \mathbf{x}_{\text{opt}}\|}{\|\mathbf{X}_0 - \mathbf{x}_{\text{opt}}\|} = \int \ln \|[\Phi]_1\|\pi(d\Phi) - \int \ln \|[\Phi]_1\|\pi(d\Phi) + \int \mathcal{R}(\Phi)\pi(d\Phi) = -\text{CR} \ .$$

The same reasoning applies for $\lim_{t\to\infty} \frac{1}{t} \ln \frac{|\lambda_t - \lambda_{\text{opt}}|}{|\lambda_0 - \lambda_{\text{opt}}|}$ and for $\lim_{t\to\infty} \frac{1}{t} \ln \frac{\sigma_t}{\sigma_0}$. Using the property of the logarithm again, we obtain

$$\lim_{t\to\infty} \frac{1}{t} \ln \frac{|\lambda_t - \lambda_{\text{opt}}|}{|\lambda_0 - \lambda_{\text{opt}}|} = \lim_{t\to\infty} \frac{1}{t} \sum_{k=0}^{t-1} \ln |\Lambda_{k+1}| - \lim_{t\to\infty} \frac{1}{t} \sum_{k=0}^{t-1} \ln |\Lambda_k| + \lim_{t\to\infty} \frac{1}{t} \sum_{k=0}^{t-1} \ln \eta^*(\varsigma * \mathbf{U}_{k+1}) \quad (39)$$

and

$$\lim_{t\to\infty} \frac{1}{t} \ln \frac{\sigma_t}{\sigma_0} = \lim_{t\to\infty} \frac{1}{t} \sum_{k=0}^{t-1} \frac{\sigma_{k+1}}{\sigma_k} = \lim_{t\to\infty} \frac{1}{t} \sum_{k=0}^{t-1} \ln \eta^*(\varsigma * \mathbf{U}_{k+1}) \ . \quad (40)$$

By applying the LLN to the right-hand sides of (39) and (40), it follows

$$\lim_{t\to\infty} \frac{1}{t} \ln \frac{|\lambda_t - \lambda_{\text{opt}}|}{|\lambda_0 - \lambda_{\text{opt}}|} = \lim_{t\to\infty} \frac{1}{t} \ln \frac{\sigma_t}{\sigma_0} = -\text{CR} \ .$$

□

6. EMPIRICAL RESULTS

By virtue of Corollary 1, all convex quadratic functions satisfy the assumptions in Theorem 1. We consider two of them in our experiments: the sphere function (f_{sphere}) and the ellipsoid function ($f_{\text{ellipsoid}}$), defined in (31) where (i) $\mathbf{H} = \mathbf{I}_{n\times n}$ for f_{sphere} and (ii) $\mathbf{H}$ is a diagonal matrix with diagonal elements $[\mathbf{H}]_{ii} = \alpha^{\frac{i-1}{n-1}}$, $i = 1, \cdots, n$, for $f_{\text{ellipsoid}}$, with condition number $\alpha = 10$. We choose $\mathrm{b} = (-1, 0, \cdots, 0)^T$ and $c = 1$ for the linear constraint $g(\mathbf{x}) = \mathrm{b}^T\mathbf{x} + c \leq 0$. According to (3) and (4), KKT conditions are satisfied for the optimum $\mathbf{x}_{\text{opt}} = (1, 0, \cdots, 0)$ and the Lagrange factor $\lambda_{\text{opt}} = 1$ for both problems.

We run Algorithm 1 and simulate the Markov chain on each problem for different parameter settings in dimensions 10, 50, and 100. We choose $k_1 = 3$, $k_2 = 5$, and $\chi = 2^{1/n}$. For space constraints, we only discuss results obtained in $n = 10$.

6.1 Single Runs

Figure 3 shows single runs of Algorithm 1 on constrained f_{sphere} (left column) and constrained $f_{\text{ellipsoid}}$ (right column) for (i) a moderate initial value of the penalty parameter $\mu_0 = 1$ (first row), (ii) a large value $\mu_0 = 10^3$ (second row), and (iii) a small value $\mu_0 = 10^{-3}$ (third row). For all runs, $\mathbf{X}_0 = (1, \cdots, 1)$, $\sigma_0 = 1$, and $\lambda_0 = 2$. Displayed are the distance to the optimum $\|\mathbf{X}_t - \mathbf{x}_{\text{opt}}\|$, the distance to the Lagrange multiplier $|\lambda_t - \lambda_{\text{opt}}|$, the penalty factor μ_t, and the step-size σ_t in log-scale, plotted against the number of iterations.

We observe that the algorithm converges linearly on both f_{sphere} and $f_{\text{ellipsoid}}$ after a certain number of iterations, independently of μ_0. The convergence on $f_{\text{ellipsoid}}$ is slower than on f_{sphere}. In the first case, the initial value $\mu_0 = 1$ is already close to the "stable" value of the penalty parameter and linear convergence of $\mathbf{X}_t$, λ_t, and σ_t towards $\mathbf{x}_{\text{opt}}$, λ_{opt}, and 0 occurs immediately. In the second case, the initial value $\mu_0 = 10^3$ is too large. However, it decreases and converges to a stable value after some iterations. The algorithm then starts to converge linearly. For a too small initial value $\mu_0 = 10^{-3}$, the distance to the optimum (and to the Lagrange multiplier) first decreases, then the algorithm stagnates. The reason is that for small values of μ_t, the Lagrange factor λ_t varies very little (see Line 6 in Algorithm 1), therefore the augmented Lagrangian does not change much between iterations, resulting in stagnation. After some iterations, however, μ_t increases again and eventually converges to a stationary value. Once μ_t is stationary, $\|\mathbf{X}_t - \mathbf{x}_{\text{opt}}\|$, $|\lambda_t - \lambda_{\text{opt}}|$, and σ_t start to decrease linearly.

6.2 Simulations of the Markov Chain

Figure 4 shows simulations of the Markov chain $(\mathbf{Y}_t, \Lambda_t, \mu_t)_{t\in\mathbb{N}}$ defined in Proposition 3 on constrained f_{sphere} (left column) and constrained $f_{\text{ellipsoid}}$ (right column), for different initial values of the penalty parameter ($\mu_0 = 1, 10^3, 10^{-3}$ in first, second, and third row respectively). The figure shows the evolution of the normalized distance to $\mathbf{x}_{\text{opt}}$, $\|\mathbf{Y}_t\|$, the normalized distance to λ_{opt}, $|\Lambda_t|$, and the penalty factor, μ_t. We choose $\mathbf{Y}_0 = (1, \cdots, 1)$ and $\Lambda_0 = 1$ in all simulations.

It can be seen from the graphs that the variables of the Markov chain seem to converge to a stationary distribution, even for too small or too large initial values of μ_t. The bump in $\|\mathbf{Y}_t\|$ and $|\Lambda_t|$ graphs we observe on the third row, for both f_{sphere} and $f_{\text{ellipsoid}}$, can be explained by looking at the third row in Figure 3: when μ_t is too small, $\|\mathbf{X}_t - \mathbf{x}_{\text{opt}}\|$ and $|\lambda_t - \lambda_{\text{opt}}|$ stagnate while the step-size σ_t decreases, resulting in an increase of $\|\mathbf{Y}_t\|$ and $|\Lambda_t|$. We observe that μ_t oscillates around about 0.1 on constrained f_{sphere} and around about 0.3 for constrained $f_{\text{ellipsoid}}$. These values are comparable to the ones we observe on single runs in Figure 3.

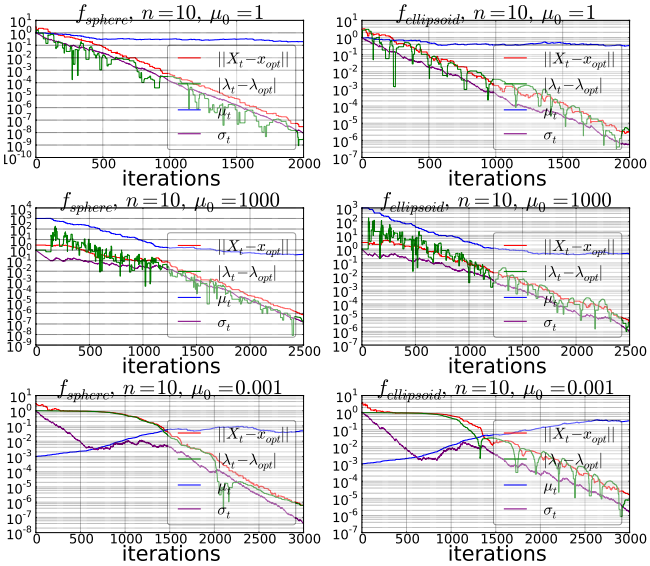

Figure 3: Single runs of the $(1+1)$-ES with augmented Lagrangian constraint handling on constrained f_{sphere} (left column) and constrained $f_{\text{ellipsoid}}$ (right column) for different initial values of μ_t in $n = 10$. Parameters of the constraint g are $\mathbf{b} = (-1, 0, \cdots, 0)^T$ and $c = 1$. $\mathbf{X}_0 = (1, \cdots, 1)^T$ and $\lambda_0 = 2$.

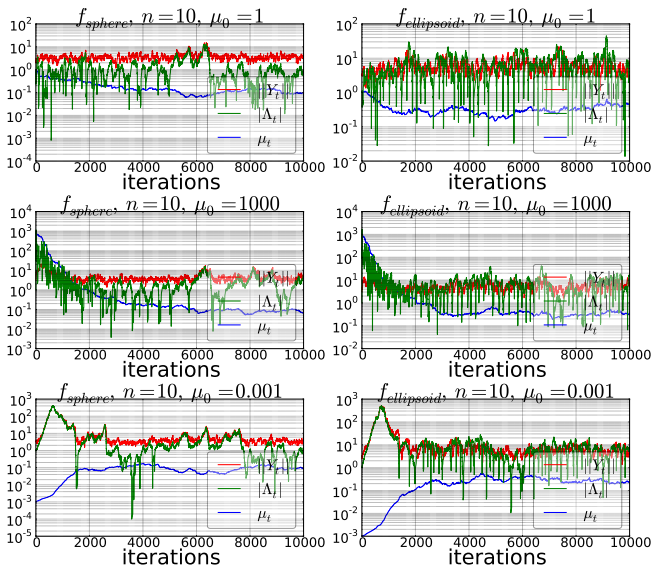

Figure 4: Simulations of the Markov chain $(\mathbf{Y}_t, \Lambda_t, \mu_t)_{t\in\mathbb{N}}$ on constrained f_{sphere} (left column) and constrained $f_{\text{ellipsoid}}$ (right column) for different initial values of μ_t in $n = 10$. Parameters of the constraint g are $\mathbf{b} = (-1, 0, \cdots, 0)^T$ and $c = 1$. $\mathbf{Y}_0 = (1, \cdots, 1)^T$ and $\Lambda_0 = 1$.

The stability of the Markov chain depends, however, on the parameters of the algorithm. In simulations not shown due to space limitations, we observe instability of the Markov chain, as well as divergence of the algorithm, for $\chi = 2$ with large values of μ_0 in $n = 100$.

7. DISCUSSION

We studied the problem of minimizing a function subject to a single linear constraint. Taking the work of [3] as a starting point, we proposed a $(1+1)$-ES with an augmented Lagrangian constraint handling approach and proved its linear convergence on problems where the associated augmented Lagrangian, minus its value at the optimum and the Lagrange multiplier, is positive homogeneous of degree 2, using a Markov chains approach, and given the stability of the considered Markov chain. To construct the Markov chain, we had to modify the update of the Lagrange factor used in [3] and consider a simpler augmented Lagrangian. Indeed, invariance alone is not sufficient, as the algorithm in [3] is translation and scale-invariant yet we could not find an underlying Markov chain.

Experiments on the constrained sphere and on the moderately ill-conditioned constrained ellipsoid showed stability of the Markov chain, as well as linear convergence of the algorithm, for the discussed parameter settings.

8. ACKNOWLEDGMENTS

This work was supported by the grant ANR-2012-MONU-0009 (NumBBO) of the French National Research Agency.

9. REFERENCES

[1] D. V. Arnold. On the Behaviour of the $(1, \lambda)$-σSA-ES for a Constrained Linear Problem. In C. A. Coello Coello et al., editors, *Parallel Problem Solving from Nature, PPSN XII*, pages 82–91. Springer, 2012.

[2] D. V. Arnold and D. Brauer. On the Behaviour of the $(1 + 1)$-ES for a Simple Constrained Problem. In G. Rudolph et al., editors, *Parallel Problem Solving from Nature, PPSN X*, pages 1–10. Springer, 2008.

[3] D. V. Arnold and J. Porter. Towards au Augmented Lagrangian Constraint Handling Approach for the $(1 + 1)$-ES. In *Genetic and Evolutionary Computation Conference*, pages 249–256. ACM Press, 2015.

[4] A. Auger. Convergence Results for the $(1, \lambda)$-SA-ES Using the Theory of φ-Irreducible Markov Chains. *Theoretical Computer Science*, 334(1-3):35–69, 2005.

[5] A. Auger and N. Hansen. Linear Convergence of Comparison-Based Step-Size Adaptive Randomized Search via Stability of Markov Chains. Submitted for publication, 2013.

[6] A. Auger and N. Hansen. Linear Convergence on Positively Homogeneous Functions of a Comparison Based Step-Size Adaptive Randomized Search: the $(1 + 1)$ ES with Generalized One-Fifth Success Rule. Submitted for publication, 2013.

[7] A. Chotard and A. Auger. Verifiable Conditions for Irreducibility, Aperiodicity and T-chain Property of a General Markov Chain. Submitted for publication, 2015.

[8] M. R. Hestenes. Multiplier and Gradient Methods. *Journal of Optimization Theory and Applications*, 4(5):303–320, 1969.

[9] S. Kern, S. D. Müller, N. Hansen, D. Büche, J. Ocenasek, and P. Koumoutsakos. Learning Probability Distributions in Continuous Evolutionary Algorithms - A Comparative Review. *Natural Computing*, 3(1):77–112, 2004.

[10] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, 1993.

[11] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, 2nd edition, 2006.

[12] M. J. D. Powell. A Method for Nonlinear Constraints in Minimization Problems. In R. Fletcher, editor, *Optimization*, pages 283–298. Academic Press, 1969.